

RNA sequencing

Paul Bertone



EBI is an Outstation of the European Molecular Biology Laboratory.

Solexa transcriptome sequencing

- Solexa data analysis and associated software development
 - Unbiased expression profiling
 - Tandem identification of expressed non-coding RNAs
 - MicroRNA identification and expression analysis
- Advantages over microarrays
 - Gene expression arrays don't capture unannotated transcripts
 - Tiling arrays are still expensive for large genomes (e.g. mammals)
 - Small RNAs are too short for stable hybridization
 - No fluorescence correction to account for, essentially zero background
- Current disadvantages
 - More expensive than standard expression arrays
 - More time consuming than any microarray technology
 - Some data analysis issues
 - No strand orientation information – sequencing a double-stranded product
 - Computing accurate transcript models, mapping reads to splice junctions
 - Contribution of high-abundance RNAs (eg ribosomal) could dilute the remaining transcript population; sequencing depth is important

Transcriptome sequencing methods

Method 1: variant of the LongSAGE protocol

Poly-A RNA selection

Double strand cDNA synthesis on beads

NlaIII digestion to remove 5' portion of cDNAs

Ligation to 5' adapters containing a MmeI recognition site

MmeI digestion to remove the 3' portion of cDNA

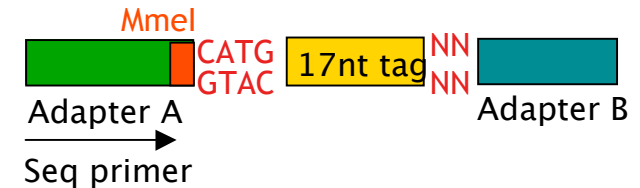
This generates a 17nt tag (not including CATG)

Tags are ligated to a 3' adapter

The construct is PCR-amplified using primers homologous to 5' and 3' adapters

PCR products are purified and quantitated (e.g. with Agilent Bioanalyzer)

Load tag-adapter hybrids into flow cell lanes and sequence

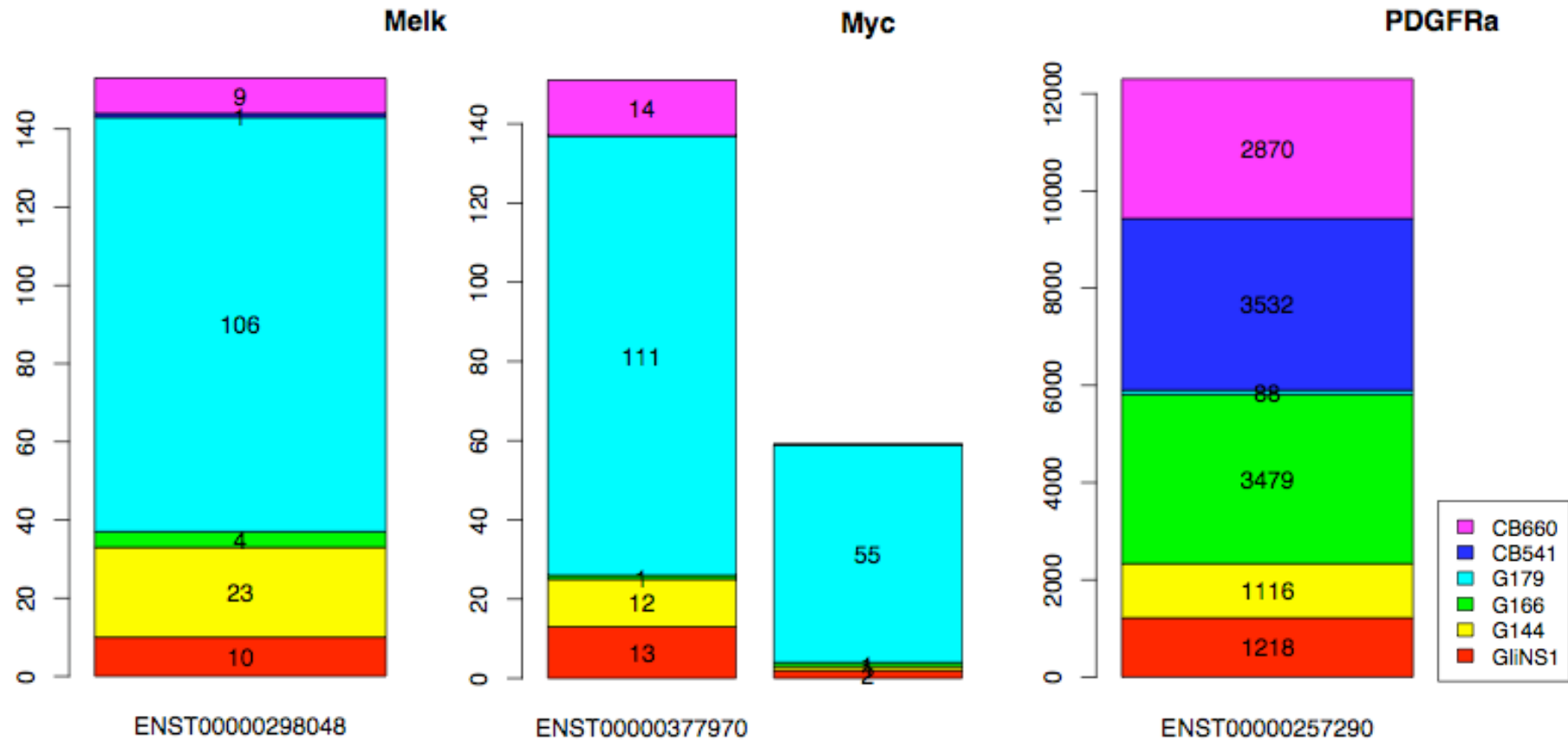


- No concatenation of SAGE tags
- One tag is amplified and sequenced per flow cell cluster
- Read (tag) alignment is performed against a library of virtual tags

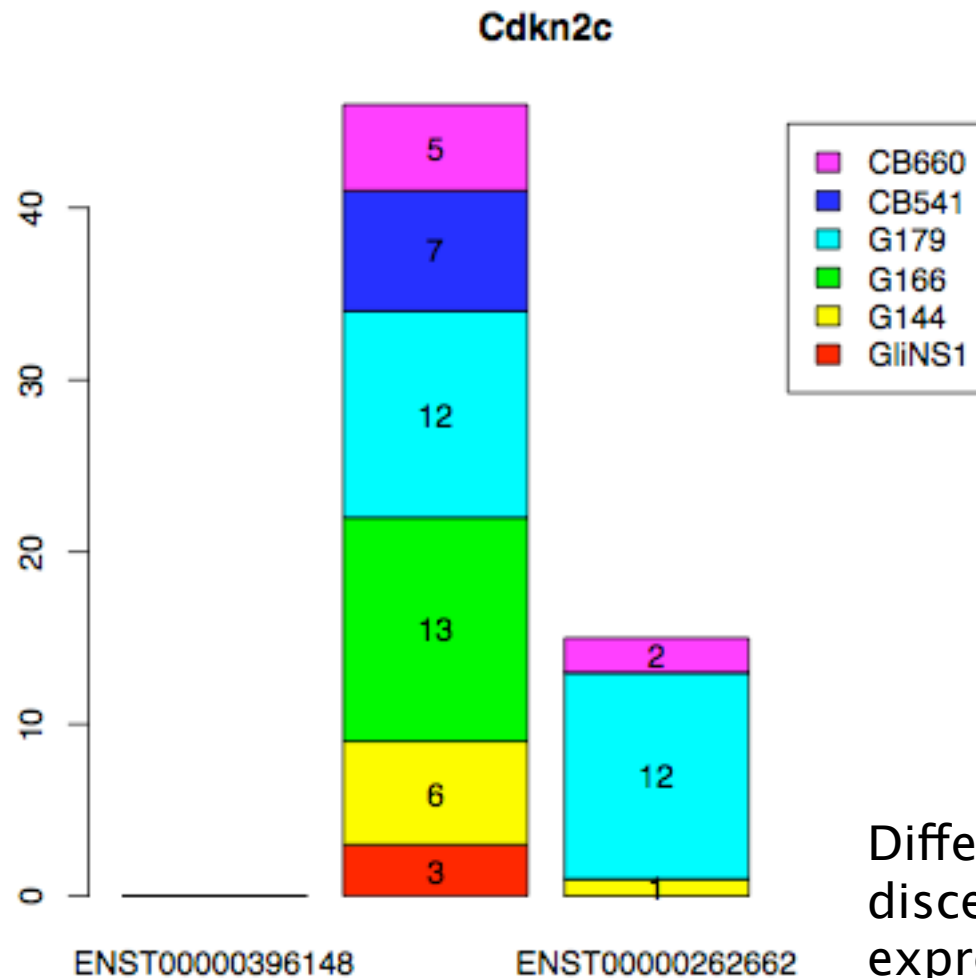
SAGE sequencing output

```
Terminal Shell Edit View Window Help
Terminal — ssh — 160x60
res-001> more FC6402_2.seq.txt
2lu: can1bt acc=676 inte739 /datCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC login: GGGGGGGGGGGGGGGG
2e or address 965 450 GTATGCCGCTCTCTGCTTGTAGTATCCGTTTTTTMicrosys: AAAAAACGGAATACTA
us: 2lu: can1bt acc=709 inte463 /datGTATGCCGCTCTCTGATGGTATATCTGTTTTTT2> c res: AAAAAACAAGATATAC
2e or address 971 634 AGTACAACAAATTTTTATCGTATCCGCTCTTTTGTecting: AACAAAAGACGGGAATAC
ge: 2lu: can1bt acc=734 inte650 /datGTATGCCGCTCTCTGCTTGTGCGTATGCCGCTCTTTTing: Per: AAAAAAGACGGCATACTA
2e or address 400 371 AAAGAAACAAAAAAAATCGTATGCCGCTCTCTCTone@res- AGAAGAAGACGGCATACTA
my 2lu: can1bt acc=665 inte473 /datGAATGCCGCTCTCTGCGTCGTATGCACTCTCTCT login: AGAAGAAGACTGCATAC
my 2e or address 57 459 GAATCCGTTTTTTTTTTCGTATGCCGCTCTCTCT: invali AGAAGAAGACGGCATACTA
my 2lu: can1bt acc=248 inte447 /datGATTTTAAATGTTTTCTTCGTATCCGCTCTCTCTTT stty: -- AAAAAAGACGGGAATAC
2e or address 345 616 AAAAAAAAAAAAAAAAAATCGGATGCCGCTCTCTTT001> dev AAAAAAGACGGCATCC
# 2lu: can1bt acc=911 inte404 /datGGGGGGCGCGTCTCTCGTTTCGTATGACGTTTTCTGT001> sol ACAGAAAACGTCATACG
my 2e or address 723 927 GTATGCCGCTCTCTGCTTTTAGTATGCCGCTCTCTTT001> ls AAGAAGACGGCATACTA
my 2lu: can1bt acc=770 inte338 /datGTTTTATATTTGTTTTTCGTATCCGCTCTTTTCTka_humar AGAAAAAGACGGGAATAC
my 2e or address 779 509 GTATGCCGCTCTCTGCTTGTAGTATGACGCTCTTTG001> dev CAAAAGACGTCATACTA
2lu: can1bt acc=778 inte365 /datGAAATCGTTTTATTTCTTCGTTTTCTCTTTTT001> cd AAAAAAGAAGGAAAAACG
# 2e or address 833 596 GTATGCCGCTCTCTGCTTGTGCGTATGCCGCTCTTT001> ls AAAAAAGACGGCATACTA
my 2lu: can1bt acc=348 inte577 /datGTATGCCGCTCTCTGCTTGTGCGTATGCCGCTCTCTT-align: AAGAAGACGGCATACTA
2e or address 659 418 GTATGCCGCTCTCTGCTTGTGCGTATCCTTTTTTTid-encode: AAAAAAAAGGAATACGA
2lu: can1bt acc=711 inte749 /datGTATGCCGCTCTCTGCTTGCAGGAAAGCCCTCTTTT31_PNB01 AAAAAAGAGGGCTTCCG
$tr 2e or address 841 246 GTATGCCGCTCTCTGCTTGTGTTTGCCTTTTTTT001> cd AAAAAAAACGGCAAACA
2lu: can1bt acc=955 inte475 /datGTATGCCGCTCTCTGCTTGTGCAATGACTCTCTTT001> ls AAAAAAGATCATTCTGA
opr 2e or address 706 473 GTATGCCGCTCTCTGCTTGTGCAAGACGCTTTT6) eland CAAAAGACGCTTACGA
2lu: can1bt acc=611 inte548 /datGTATGCCGCTCTCTGCTTGTGCTAAATCTCTTTTT001> ls AAAAAAAAGAATTAGA
wh 2e or address 292 518 GGTTCATAAAAGTTTTTCGTATTTCTCTTTTGT31_PNB01 AGCAAAACAAGAAATAC
2lu: can1bt acc=37 inte325 /datAGAAAAAAAAAAAAAAAAATCGTATGCCGCTCTCTGCT001> ls AGCAGAAGACGGCATACTA
otl 2e or address 582 759 AACATCAAACTTTTGTTTCGTATGCCGCTCTCTGCTeland/_l AGCAGAAGACGGCATACTA
2lu: can1bt acc=612 inte364 /datGTATGCCGCTCTCTGCTTGGCGTATCGTCTCTCT001> ls AGAAGAAGACGATACGC
2e or address 988 773 GTATGCCGCTCTCTGCTTGTAGTATGACGATTTTT001> bjc AAAAAATACGTCATACTA
te 2lu: can1bt acc=971 inte650 /datATAAAAGGAATCAGAATTTTCGTATGCCGCTCTCTGChfinishe GCAGAAGACGGCATACTA
2e or address 588 543 GTATGCCGCTCTCTGCTTTAGAATGCCGCTCTTT001> ls AAAAAAGACGGCATTCTA
on 2lu: can1bt acc=971 inte473 /datGTTCTTACTGAGATTTTCGTATGCCGCTCTCTGCT 846464 AGCAGAAGACGGCATACTA
} 2e or address 605 367 ATAATAAAAGCTATTATTCGTATGCCGCTCTTTTTTrw-r-- AAAAAAGACGGCATACTA
2lu: can1bt acc=943 inte709 /datGAATTCGCTCTCTGCTTTCGTATGCTGTTTTTTTrwxr-x AAAAAAAACAGCATACG
```

Solexa transcriptome sequencing



Solexa transcriptome sequencing



Differential read counts allow us to discern which transcript isoforms are expressed

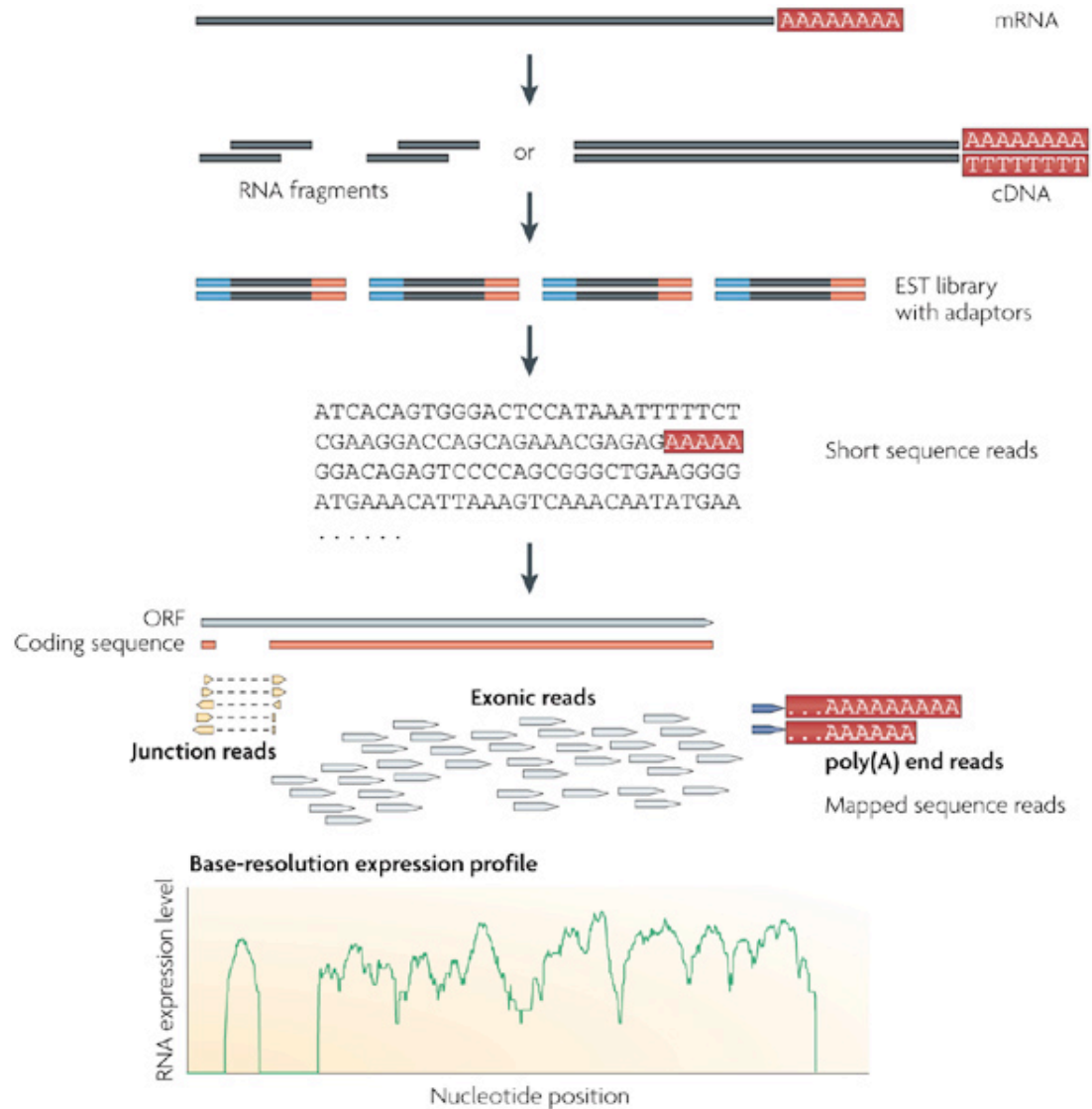
Features of SAGE analysis

- Complicated library construction
- Good at gene expression analysis
- Short reads (17nt), therefore low rate of unique alignments to reference genome
 - Reads are mapped to virtual tags instead
- Mostly limited to annotated genes
- Can get some information on novel transcripts (limited)

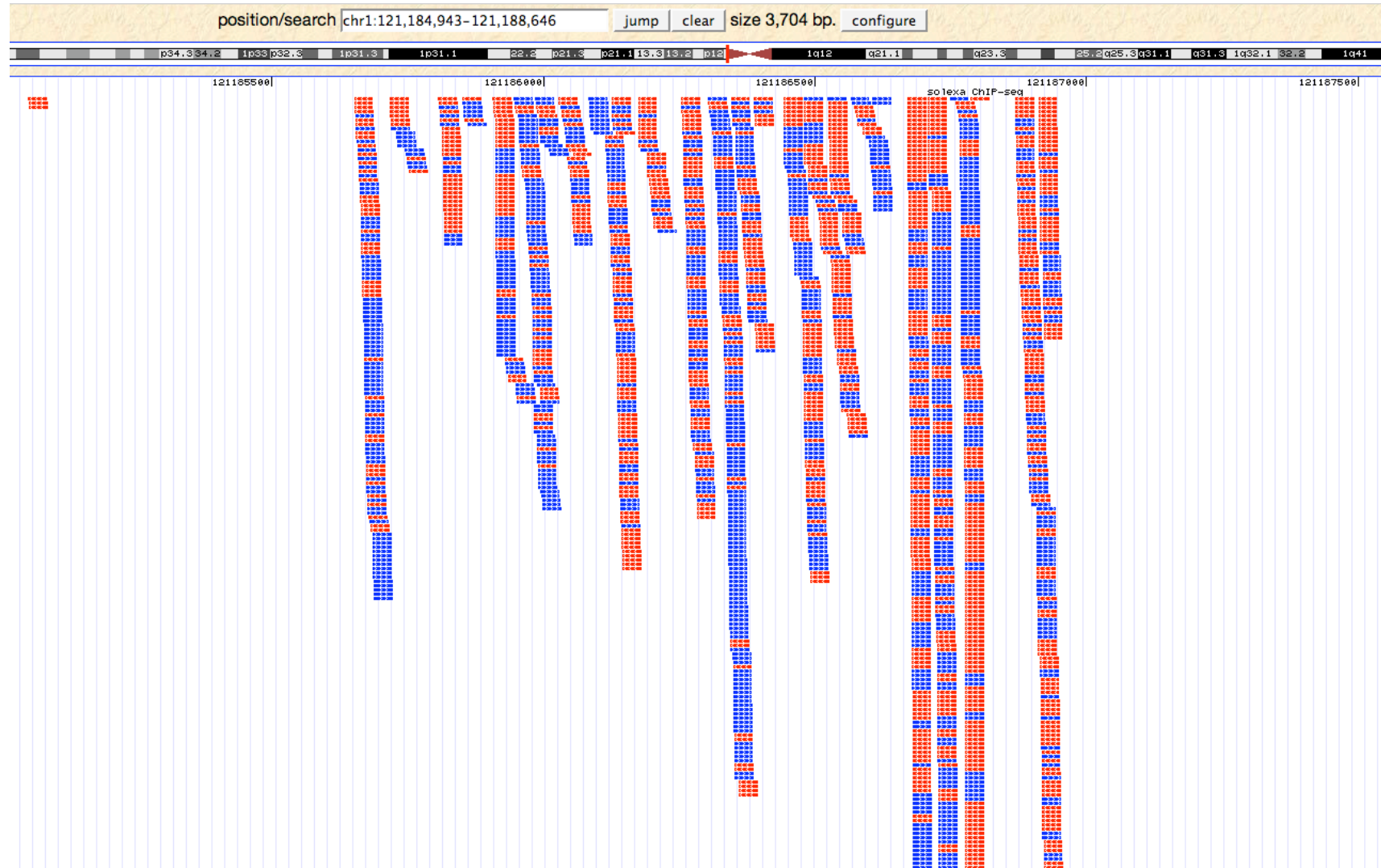
mRNA sequencing

- Similar to SAGE analysis in terms of gene expression
- Simpler library construction
- Not limited to 17nt reads
 - Utilize full read length for alignment
 - Much better genome mapping
- Results are analogous to tiling array profiling
 - Reads map to individual transcript components
 - Ascertain splice variation as well as gene expression
 - Refine existing annotation of exons and UTRs
 - Identify non-coding RNAs

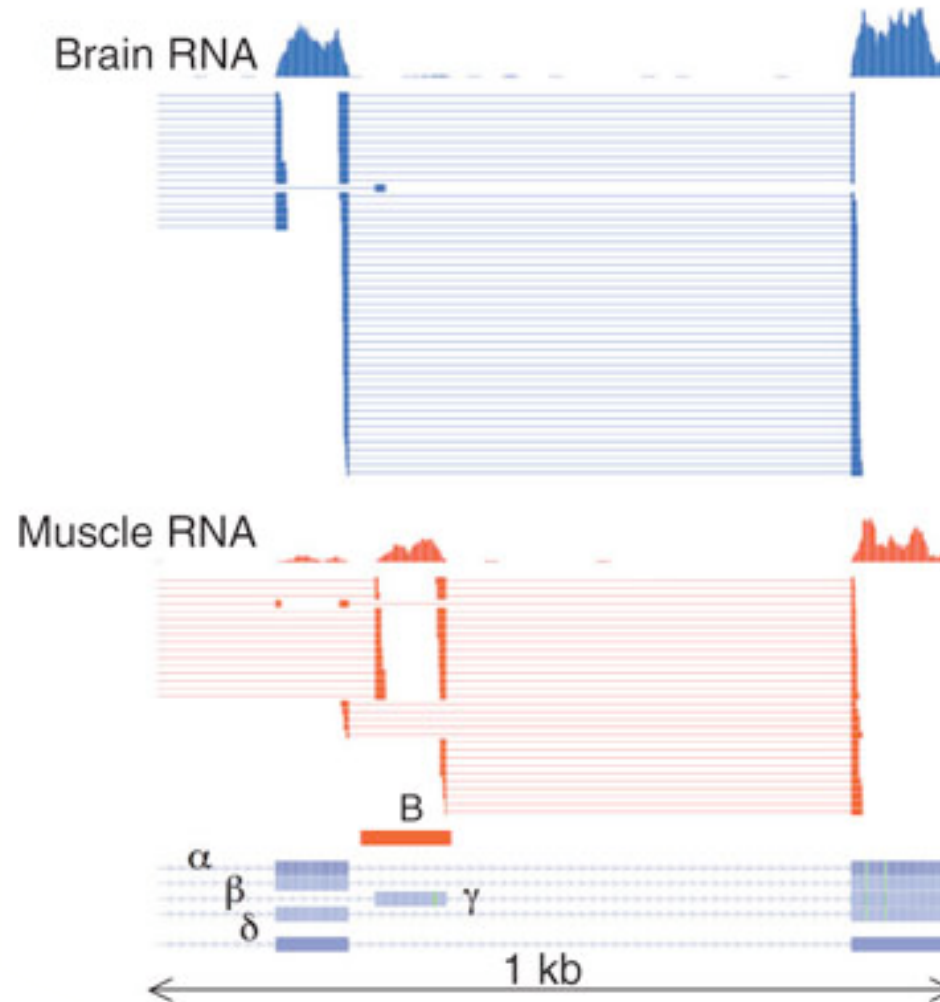
mRNA-seq protocol



Reads mapped to the human genome

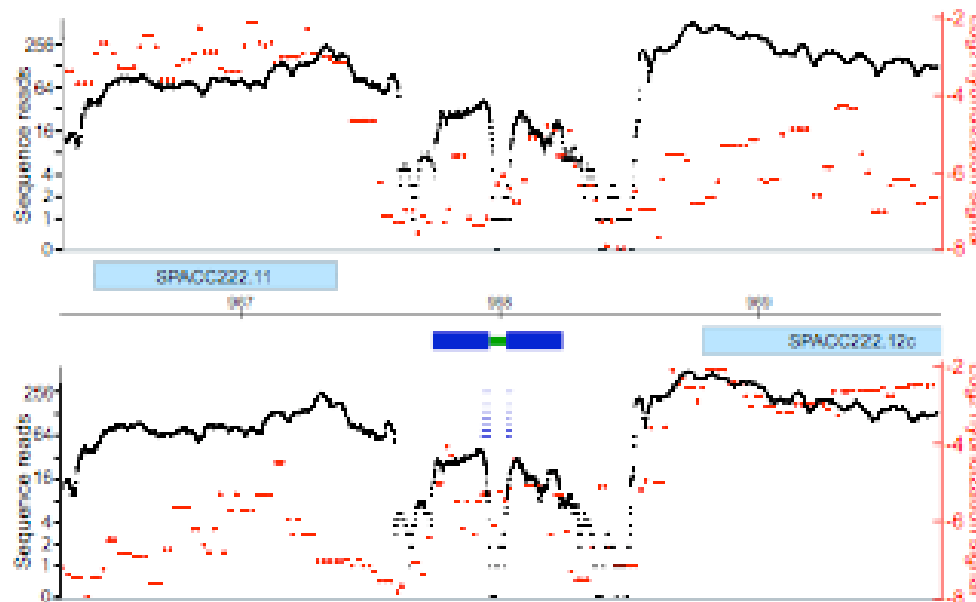


Alignment to exon splice junctions



Alignment reference should consider mature transcripts or exon junctions

Sequencing vs. tiling array hybridization



Red: tiling array hybridization signal (log₂)
Black: sequencing reads (log)

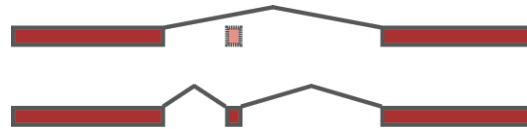
- Example comparison between Solexa WTSS and tiling array hybridization data (S. pombe, Bahler lab Sanger)
- Top image = sense strand; bottom image = antisense strand
- Light blue = annotated genes; Dark blue = new non-coding transcript; Green = intron

Novel Transcribed Regions: Possibilities

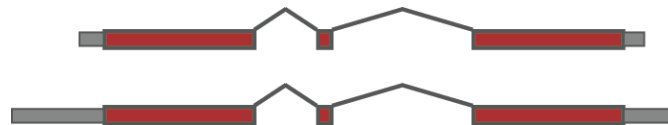
- Many areas of active transcription are observed outside annotated genes
 - Rare or low-abundance protein-coding transcripts



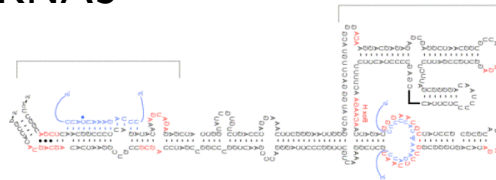
- Unannotated exons from alternate splice products



- Previously under-represented 3' and 5' UTRs



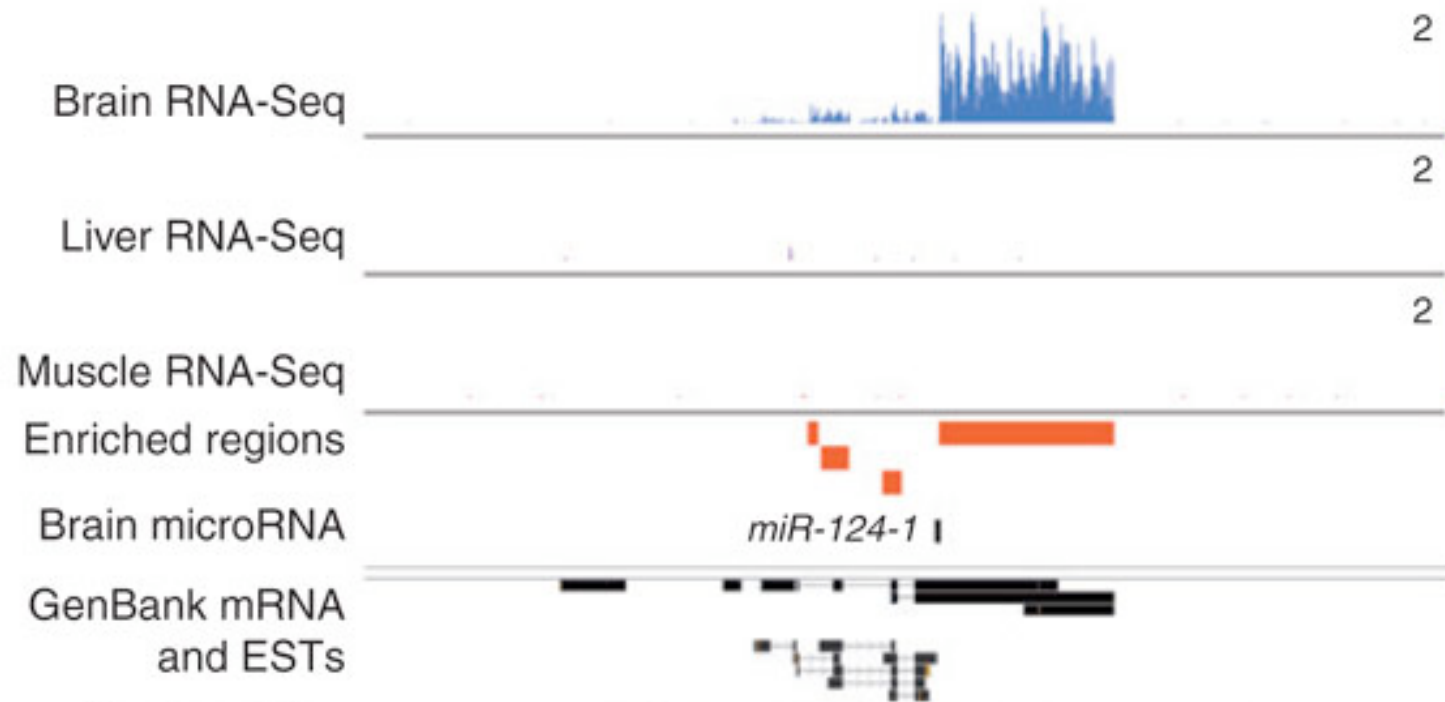
- Noncoding RNAs



Splice variation, refinement of existing exon annotation



Detection of microRNA precursors



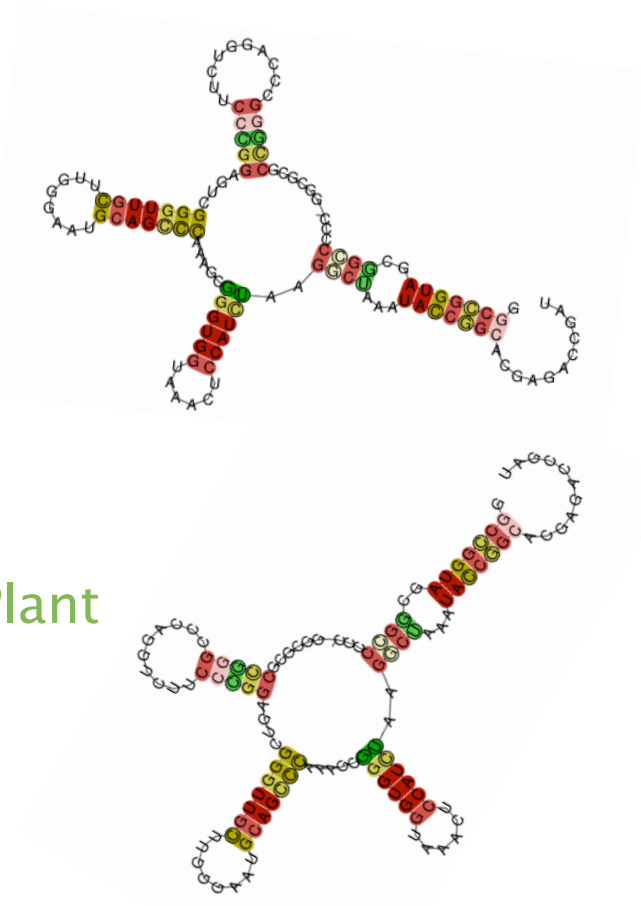
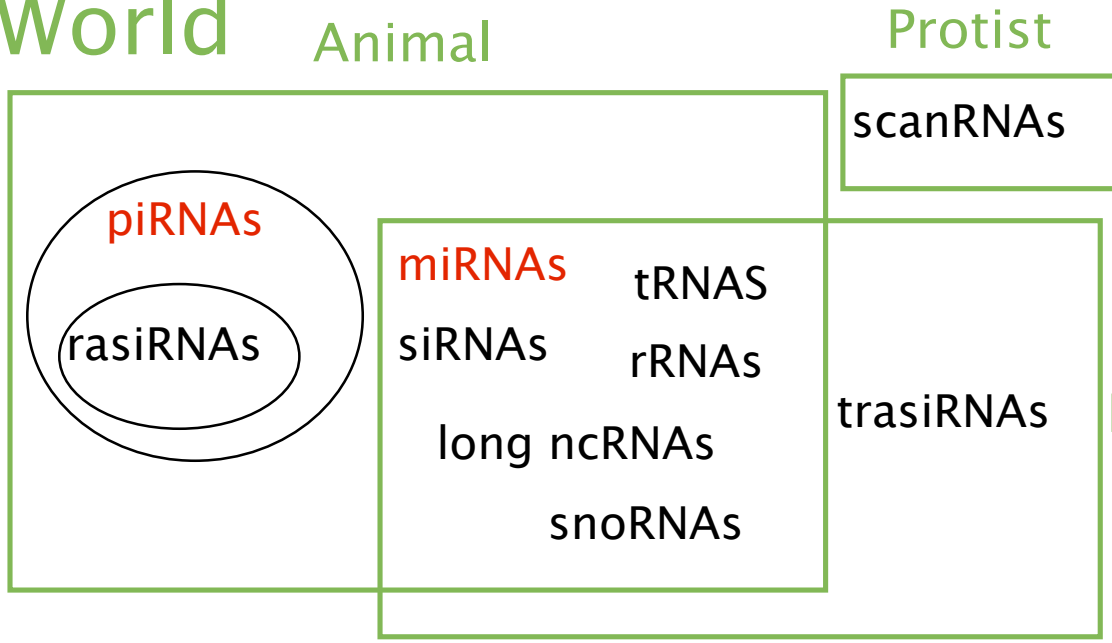
Protocol variations

- Fragmentation methods
 - RNA: nebulization, hydrolysis
 - cDNA: sonication, Dnase I treatment
- Depletion of highly abundant transcripts
 - e.g. RiboMinus – others?
- Oligo-dT selection for poly(A)+ transcripts vs total RNA
- Coverage issues
 - What is the sequencing depth required?
- Strand specificity
 - Most RNA sequencing is not strand-specific
 - Currently working with Vladimir Benes and Lars Steinmetz on new protocols for this

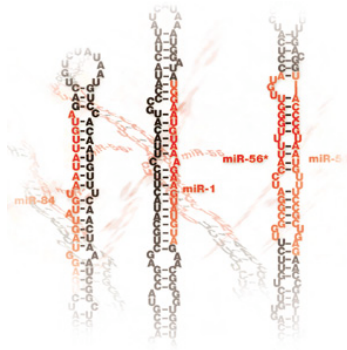
Specialized RNA-seq applications

- Small RNA sequencing
 - microRNAs
 - piRNAs
 - endo-siRNAs
- Identification of RNAs associated with protein complexes (e.g. Ago2)
 - Immunoprecipitation of RNA-bound protein complexes
 - Proteinase K digestion, purification of nucleic acids for sequencing

The Non-coding RNA World



- Growing number of non-coding RNA classes categorized by many different features (e.g. function, length, secondary structures, expression tissues, species, etc.)
- For my projects I am focusing on short regulatory non-coding RNAs..
- ..paying particular attention to the microRNA and piwiRNA classes



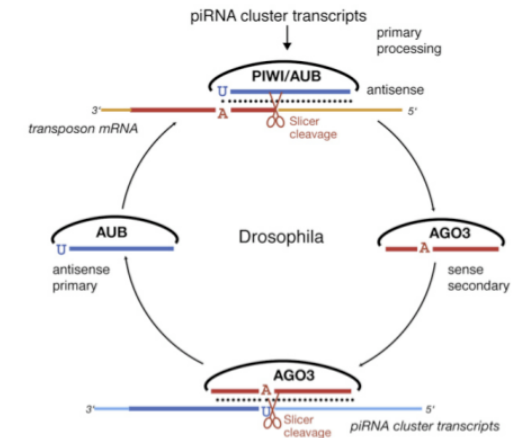
microRNAs and piwiRNAs

Differences

- miRNAs are generally shorter (~21–23nt) than piRNAs (~24–30nt)
- miRNAs are Dicer-dependent
- miRNAs are processed from a dsRNA precursor with a known secondary structure (piRNAs?)
- Expression of piRNAs is thought to be restricted to the germline
- miRNAs bind to Argonaute clade whilst piRNAs to the Piwi clade of the Argonaute protein family

Similarities

- Both show a 5'Up preference
- Both show a 2'O-methyl modification at their 3' end (plant microRNAs only)



Differences in small RNA sequencing

- Size exclusion of total RNA
 - Selected to target particular species
 - e.g. 17–23nt for microRNAs, 25–32nt for piRNAs
 - 17–32nt can encompass both populations
- Direct ligation of adapters to RNA molecules
- Transcripts are typically shorter than the reads
 - Sequence into the adapters
 - Reveals strand specificity

Adapter masking, low-complexity filtering

```
@HWI-EAS225_30EK7AAXX:6:1:1481:96      @HWI-EAS225_30EK7AAXX:6:1:1481:96      >128
GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATTATA  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATTATA  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATTATA
+HWI-EAS225_30EK7AAXX:6:1:1481:96      +HWI-EAS225_30EK7AAXX:6:1:1481:96      >129
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA] ^NNNNH<-----  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA] ^NNNNH<-----  >129
GTTAATGTATCTATGGACTTAAAAATGGCANNNNNNN  GTTAATGTATCTATGGACTTAAAAATGGCANNNNNNN  >130
GGAAATGATGAGCCAGAAGATTCAACAGCANNNNNNN  GGAAATGATGAGCCAGAAGATTCAACAGCANNNNNNN  >131
>130
+HWI-EAS225_30EK7AAXX:6:1:1668:1848    +HWI-EAS225_30EK7AAXX:6:1:1668:1848    +HWI-EAS225_30EK7AAXX:6:1:1668:1848
GTTAATGTATCTATGGACTTAAAAATGGCANNNNNNN  GTTAATGTATCTATGGACTTAAAAATGGCANNNNNNN  GTTTTCTAGGAAAAGTTTTGGCTGTTGTATGNNNN
+HWI-EAS225_30EK7AAXX:6:1:1668:1848    +HWI-EAS225_30EK7AAXX:6:1:1668:1848    +HWI-EAS225_30EK7AAXX:6:1:1668:1848
AAAAAAAAAAAAAAAA [AAAAAAAA^NNNNN<-----  AAAAAAAAAAAAAAAAA [AAAAAAAA^NNNNN; ; ; ; ;  >132
>131
@HWI-EAS225_30EK7AAXX:6:1:1548:1360    @HWI-EAS225_30EK7AAXX:6:1:1548:1360    @HWI-EAS225_30EK7AAXX:6:1:1548:1360
GGAAATGATGAGCCAGAAGATTCAACAGCCTCGTATG  GGAAATGATGAGCCAGAAGATTCAACAGCANNNNNNN  GGAAATGATGAGCCAGAAGATTCAACAGCANNNNNNN
+HWI-EAS225_30EK7AAXX:6:1:1548:1360    +HWI-EAS225_30EK7AAXX:6:1:1548:1360    +HWI-EAS225_30EK7AAXX:6:1:1548:1360
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA^YFNNNNN<-----  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA^YFNNNNN; ; ; ; ;  >133
>132
@HWI-EAS225_30EK7AAXX:6:1:1278:1293    @HWI-EAS225_30EK7AAXX:6:1:1278:1293    @HWI-EAS225_30EK7AAXX:6:1:1278:1293
GTGTTCCCTAGGAAAAGTTTTGGCTGTTGTATGTCGT  GTGTTCCCTAGGAAAAGTTTTGGCTGTTGTATGNNNN  GTGTTCCCTAGGAAAAGTTTTGGCTGTTGTATGNNNN
+HWI-EAS225_30EK7AAXX:6:1:1278:1293    +HWI-EAS225_30EK7AAXX:6:1:1278:1293    +HWI-EAS225_30EK7AAXX:6:1:1278:1293
AAAAAAAA [AZ^Y^AAAAAAAA^T^F^HNNN<-----  AAAAAAAAA [AZ^Y^AAAAAAAA^T^F^HNNN<; ; ; ;  >134
>133
@HWI-EAS225_30EK7AAXX:6:1:177:227      @HWI-EAS225_30EK7AAXX:6:1:177:227      @HWI-EAS225_30EK7AAXX:6:1:177:227
GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATAATA  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATAATA  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATAATA
+HWI-EAS225_30EK7AAXX:6:1:177:227      +HWI-EAS225_30EK7AAXX:6:1:177:227      +HWI-EAS225_30EK7AAXX:6:1:177:227
AAAAAAAA [AAAAAAAAAAAAAAAA^NNNNN<-----  AAAAAAAAA [AAAAAAAAAAAAAAAA^NNNNN<-----  >135
>134
@HWI-EAS225_30EK7AAXX:6:1:47:1634      @HWI-EAS225_30EK7AAXX:6:1:47:1634      @HWI-EAS225_30EK7AAXX:6:1:47:1634
GAACAGATGGCTTCCCACATGTACAGTCGTATGCCG  GAACAGATGGCTTCCCACATGTACAGNNNNNNNNNN  GAACAGATGGCTTCCCACATGTACAGNNNNNNNNNN
+HWI-EAS225_30EK7AAXX:6:1:47:1634      +HWI-EAS225_30EK7AAXX:6:1:47:1634      +HWI-EAS225_30EK7AAXX:6:1:47:1634
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA^NNNNN<-----  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA^N; ; ; ; ; ; ; ;  >136
>135
@HWI-EAS225_30EK7AAXX:6:1:1099:113     @HWI-EAS225_30EK7AAXX:6:1:1099:113     @HWI-EAS225_30EK7AAXX:6:1:1099:113
GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATCTGTT  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATCTGTT  GTATGCCGTCTTCTGCTTGAAAAAAAAAAAAATCTGTT
+HWI-EAS225_30EK7AAXX:6:1:1099:113     +HWI-EAS225_30EK7AAXX:6:1:1099:113     +HWI-EAS225_30EK7AAXX:6:1:1099:113
AAAAAAAAAAAAAAAAAAAAAAAA^Y^TNSNNNNN<-----  AAAAAAAAAAAAAAAAAAAAAAAAA^Y^TNSNNNNN<-----  >137
>136
@HWI-EAS225_30EK7AAXX:6:1:1561:621     @HWI-EAS225_30EK7AAXX:6:1:1561:621     @HWI-EAS225_30EK7AAXX:6:1:1561:621
GAGGAAAGTAGACTCTCAGAACAAGTCGTATGCC  GAGGAAAGTAGACTCTCAGAACAAGNNNNNNNNNN  GAGGAAAGTAGACTCTCAGAACAAGNNNNNNNNNN
+HWI-EAS225_30EK7AAXX:6:1:1561:621     +HWI-EAS225_30EK7AAXX:6:1:1561:621     +HWI-EAS225_30EK7AAXX:6:1:1561:621
AAAAAAAAAAAAAAAAAAAAAAAA^ \^NNNNN<-----  AAAAAAAAAAAAAAAAAAAAAAAAA^ \^NN; ; ; ; ; ; ; ;  >138
>137
>138
>139
>140
>141
>142
>143
GAACAGGACACAGAAGGAGCTGTTTCATANNNNNNN  GAACAGGACACAGAAGGAGCTGTTTCATANNNNNNN  GAACAGGACACAGAAGGAGCTGTTTCATANNNNNNN
```

Aligned RNA reads from RNA-seq

```

chr10 18519329 G 0 @
chr10 18519330 C 0 @
chr10 18519331 C 14 @ggaa.a..g.gggd
chr10 18519332 A 27 @.gg.c.....cc.gggc..g...g
chr10 18519333 T 27 @aaaagaggagaaaaaadaaaaaa
chr10 18519334 G 27 @a.aaa.aa.a....aaaa.aaaaa.a
chr10 18519335 T 27 @gaggaaaaaadaaaaggagagggag
chr10 18519336 A 27 @.c.g.gggccc.....
chr10 18519337 G 27 @c.cca.aa.a....ccc.caccac.c
chr10 18519338 G 27 @...cacc.c.aaa..a.....
chr10 18519339 T 27 @gcgggcgggcgccccggggcggggcg
chr10 18519340 T 27 @caccggggagagggcccccaccaccac
chr10 18519341 T 27 @a.aacgcc.c..gggaagaaaaa.a
chr10 18519342 C 27 @tatta.aaaaa...ttt.tttttat
chr10 18519343 A 27 @.c.t.tttcc.....e.
chr10 18519344 G 27 @c.ccataa.a.tttccctcccc.c
chr10 18519345 A 27 @g.ggc.cc.c....ggg.g.ggg.g
chr10 18519346 A 27 @.g..gcgggggccc...c.....g.
chr10 18519347 A 27 @ttt.g..t.tgggttgtttttttt
chr10 18519348 A 27 @.....
chr10 18519349 A 27 @ggggggggggggggggggggggggg
chr10 18519350 C 27 @.....
chr10 18519351 A 27 @c.ccccccccccccccccccccccc
chr10 18519352 G 27 @aaaaaaaaaaaaaaaaaaaaaaaaa
chr10 18519353 T 27 @.....
chr10 18519354 G 27 @.....
chr10 18519355 A 27 @ttttttttttttttttttttttttt
chr10 18519356 A 27 @.....
chr10 18519357 A 27 @ggggggggggggggggggggggggg
chr10 18519358 G 27 @.....
chr10 18519359 C 27 @ttttttttttttttttttttttttt
chr10 18519360 C 27 @ttttttttttttttttttttttttt
chr10 18519361 C 27 @ttttttttttttttttttttttttt
chr10 18519362 T 27 @cccccccccccccccccccccccc
chr10 18519363 T 27 @aaaaaaaaaaaaaaaaaaaaaaaaa
chr10 18519364 C 27 @ggggggggggggggggggggggggg
chr10 18519365 A 27 @.....
chr10 18519366 A 27 @.....
chr10 18519367 G 27 @aaaaaaaaaaaaaaaaaaaaaaaaa
chr10 18519368 G 27 @aaaaaaaaaaaaaaaaaaaaaaaaa
chr10 18519369 A 27 @.....
chr10 18519370 T 27 @cccccccccccccccccccccccc
chr10 18519371 G 33 @aaaaaaaaaaaaaaaaaaaaaaaaa
chr10 18519372 C 33 @ggggggggggggggggggggggggg
chr10 18519373 C 33 @ttttttttttttttttttttttttt
chr10 18519374 A 33 @ggggggggggggggggggggggggg
chr10 18519375 A 33 @.....cc.c.
chr10 18519376 T 19 @aaaaaaaaaaaaaggggggg
chr10 18519377 A 6 @.gg.g.
chr10 18519378 C 6 @.a...
chr10 18519379 C 6 @gaagag
chr10 18519380 C 6 @gttgtg
chr10 18519381 A 6 @c..c.c
chr10 18519382 G 6 @accaca
chr10 18519383 T 6 @.aa.a

```

Sample	KS35	KS45
Reads	3,559,384	5,861,316
Eland placement (total)	2,429,078 (68%)	4,109,776 (70%)
Unique, no mismatch	1,806,384	3,445,856
Unique, 1 mismatch	418,151	440,111
Unique, 2 mismatches	204,543	223,809

Transcriptional units from RNA-seq

```
>MM9:10:18509523-18509567
AGTAGGAAGTTATGGTATCTTTGGAAAGTCAGTTGTGTTAGCTGG
>MM9:10:18516039-18516083
TGGTGGAGTCTTCTTTTGGTTGTCATTGGGGACCTATAGAGGGCA
>MM9:10:18516279-18516359
ATGTGCTTAATTTTGTATTCTGTATACCTTCTGGCATTGCTGATCACCTACATTCATTATTATTATTACC
>MM9:10:18516802-18516857
ACTTCTAGTCTACCTATGTCCATCTAACCTGCCTTCTTCCCACTTCAGAGAT
>MM9:10:18516888-18516932
ATGACCCCTCTAGCTGCTTCATGCTGTGAGGGGGCGGAGTCCAGG
>MM9:10:18516988-18517060
CTTCTTTAGACAGGCTGTACGTATCTGAGGCTCCACGTAAGGGCCATCTGTAGAAGTGGAAAGGGCTGAAGGA
>MM9:10:18517358-18517407
TATCTAGGCAGCATTCTTCTTATCAAGAAGAAAACTACTAATGAGAAATG
>MM9:10:18517447-18517515
GAGTATGGTCCCTTAAACTGTTCAAGAAATACACCAGTCATTTATTTATTTATTTTGGTTATA
>MM9:10:18517618-18517724
TGAGCGCTAACCTTCTTTTCCAAGTATGCTACTTACGGGAGGAGAATTTAGCTTAGCTTAGTGGCTTACTTGGAAACATGGAAGCAAGATCCATGGAGGGTCC
>MM9:10:18517761-18517826
TTGAGTATCTGGTTGGCGTCTCCATGACCTCCCTGAACAACAGAAAAGTGATTCCTCACAGTTA
>MM9:10:18517836-18517888
TTGTGTGAGATGCTTGTCCGACCTACTTGATCTTGGGGCCCAAGGAGGAATAT
>MM9:10:18517895-18517964
CTTTTAAATCCCTCTGACCACATTAGTATGGTCTCCAAGTATGGTTATTGAACACCCAGGATGCCACTGA
>MM9:10:18518002-18518149
TTTTTTTCTCAAAACAATCTTTGTAATTGTTAGGGAGAACAGGCCATTTATTTAAGCAGAGCTTTGACATCAATGTATTGAATCTGACTGTTTACATACCTTATAAATCTGCTCAGACCCGTATAATTGCTTAAGCTTTCCAAT
>MM9:10:18518187-18518268
GAGACATAATAACATTAACAAGCCAATAATCAACAAGACTAGCAGTGATGATCCTTTGAAATTTGGCATAGCAATCCTTGGC
>MM9:10:18518350-18518399
TAAATTACATGTATGCATCTGTTAAATTAATGGGCCCTGTTAAGTTTCCA
>MM9:10:18518496-18518549
AGAACAAAAGTGGTCAATAGTCCCAAGATCATCAGGGGAGCAGGATCTATGGCA
>MM9:10:18518610-18518711
ATCTGTTGGTCTGGGAGCTTCTTAAAGTCCCAAGGAGCTTATGCACTACAGTAATCTTCAAGCATCTATCAAGCTTCAAGAACCAGCTCAATTAACCTCT
>MM9:10:18518781-18518865
ATTTTAAAAAGTACATATATATATACACATACACACTAGGTGCTTCTTTTAAAGATCCTTGGCCCAAGAGATCACAATCTATG
>MM9:10:18518930-18519120
ACACCTTCTAGATCTTGGCTTCTCTCCAAAGTGCCCTCTACAGAGTTAAATAGACATGGGTATGTACTTGTAGCCTCTCTGACTCACTAGCAACCATGGAAGAAGGGGAGCTTGTTCATAGCATTGGCTCAAGGCAATGCCAAAGTCTCCCG
TAAGCCATCAGAGAGCTCTTACACACAAC
>MM9:10:18519161-18519278
TGGTAGAGGTACCTGCAATGTCAGAAGTGTGATGCACTCCAAAGTCCAGTTAAGGGAATCTGTTCTGCCACCACATCTTGTCTCACATATGATAGCAACTTTTGGAGCCAGGACA
>MM9:10:18519282-18519327
ATACATCTCTTGAGAGCCCTAAGTAAGAGCAAGACAGACCTCCAGT
>MM9:10:18519331-18519521
CATGTAGGTTTCAGAAAAACAGTGAAGCCCTCAAGGATGCCAATCCCACTACATCTTCTTATGCACAGAAAAGACCTTACCCTATAGCTCACTCATAGCCGTCACCAATTTGAACTTCAAGCAACTTCAATTTGAACTCAATGGAAGCTTGTCTGGT
GGTGCCAAAGAGTGCAATGCCACCTGACTTA
>MM9:10:18519523-18519641
TTCAGGCTGCTGGCCACTTTGCTTAAGCAGAGTCTCTTACTGGCCTATAGTTTCTAAGTAGATTAGGCTAGCCATCCGATGAGTCCCAAGGAATACACACCTGTTGGCTCACCTTGGCAC
>MM9:10:18519649-18519693
GGTGTCTTTCACTGTACTTCCCGCCTCACACTTGCATGGAGG
>MM9:10:18519725-18519969
TATGGTTTACTTCAATCGTCTTTCATCACTAGGTTCCCTGGTCTTTTCCACATTTGGCTTTGTAGATTATCTTCAATCAATTTTCAATTTTGTGTTGAGACTTTTGTCAAATTTTAAAGGTTATATTTTGGTATTTCTTTTACTTTCTTTCAGTC
TCTTTTTTCTTGTAAATATTTCTTGGCTCTAGTAATACTCTGGATCTGTGGAATTTACTACAATTTGGCATTAGTAGG
```


Transcriptional units from RNA-seq

```
chr10 18522140 G 6 @A....  
chr10 18522141 G 6 @Accccc  
chr10 18522142 A 6 @Tcccc  
chr10 18522143 G 6 @Cttttt  
chr10 18522144 G 6 @,tttt  
chr10 18522145 T 13 @,....gg.a.a.  
chr10 18522146 A 13 @,cccccccccc  
chr10 18522147 G 13 @Tccccaaaaaa  
chr10 18522148 C 13 @Gaaaaagagaga  
chr10 18522149 A 13 @Cggggg.g.g.g  
chr10 18522150 A 18 @C.....c.c.c.....  
chr10 18522151 A 18 @G.....gg.ggg.ggggg  
chr10 18522152 T 18 @,ccccccagagaaaa  
chr10 18522153 G 17 @aaaaaccacaccccc  
chr10 18522154 G 17 @....t.a.a.a.....  
chr10 18522155 C 17 @aaaaatgggggggggg  
chr10 18522156 A 17 @....cc.c.c.ccccc  
chr10 18522157 G 12 @aaccccaaaaa  
chr10 18522158 A 12 @ttgtgtgtttt  
chr10 18522159 G 12 @ta.a.a.aaaa  
chr10 18522160 G 12 @cccccccccccc  
chr10 18522161 T 12 @ggagagaggggg  
chr10 18522162 T 12 @aa.a.a.aaaa  
chr10 18522163 C 19 @ggagagagaaaaaaa.aaa  
chr10 18522164 T 19 @,ggcgcggggggaaagg  
chr10 18522165 T 19 @,g.g.g.g....ccgggcc  
chr10 18522166 C 19 @aaaaaaaaaaaaa.a.aa  
chr10 18522167 A 19 @ggggggggggggg...gg  
chr10 18522168 T 19 @cccccccccccaaggga  
chr10 18522169 A 19 @.....  
chr10 18522170 G 19 @aaaaaaaaaaaaa..aca..  
chr10 18522171 G 19 @aaaaaaaaaaaaaa...aa  
chr10 18522172 G 19 @,tttttttttcca.acc  
chr10 18522173 T 19 @ggggggggggggggccgg  
chr10 18522174 A 19 @ggggggggggggggg.ggg  
chr10 18522175 T 19 @ccccccccccccccgg.gcc  
chr10 18522176 T 19 @aaaaaagaaaaaacacaa  
chr10 18522177 C 19 @ggggggggggggtta.att  
chr10 18522178 T 19 @aaaaaagaaaaa.g.aa  
chr10 18522179 A 19 @ggggggggggggc...cc  
chr10 18522180 G 19 @.....c.c..  
chr10 18522181 T 19 @,.....aag.gaa  
chr10 18522182 C 19 @gttttttttttatatt  
chr10 18522183 A 19 @cccccccccccccccccc  
chr10 18522184 T 19 @.....  
chr10 18522185 C 19 @tttttttttttttttt  
chr10 18522186 T 19 @cccccccccccccccccc  
chr10 18522187 A 19 @.....  
chr10 18522188 T 19 @.....  
chr10 18522189 A 19 @.....  
chr10 18522190 C 12 @gggggggggggg  
chr10 18522191 C 12 @gggggggggggg  
chr10 18522192 A 12 @gggggggggggg  
chr10 18522193 T 12 @.....  
chr10 18522194 G 12 @aaaaaaaaaaaa  
chr10 18522195 G 7 @tttttt  
chr10 18522196 C 7 @tttttt  
chr10 18522197 A 7 @cccccc  
chr10 18522198 G 7 @tttttt
```

...

```
chr10 18522085 G 0 @  
chr10 18522086 T 51 @,aa.g....caaaa.aaaa,aca,a,gaga,aa,d,acag,.d,aaag  
chr10 18522087 G 51 @cc,.cccc,c.a.c.c....c.a.c.ta.a.c.c.c.a.occ.caa.d  
chr10 18522088 G 51 @aacaaaaaaa,aaaaaaccccccaaacacaaaaaac,aaaaaaaca  
chr10 18522089 C 51 @aaaaaaaaaaaaaaaaaaaaaaag,aaagaaaaaaaagaaaaag  
chr10 18522090 T 51 @gggggggggggagaggggggggggggaggggggggaggggggga  
chr10 18522091 C 51 @,.aa,....aga,.d,aaaa,aag,ag,d,a,aa,d,aga...d,  
chr10 18522092 C 51 @aaa.agaaaa,a.g,a,d,aaa,g,d,.g,aaa,.d,aaa,aaa,agaag  
chr10 18522093 T 51 @ggggggggggggcgggggggggggggggggggggggggggggggg  
chr10 18522094 A 51 @,.g.c...ggg,.g,d,g,.g,t,gggc,.gg,g,.ggc,.g.c.c.  
chr10 18522095 T 51 @aaaaaaaaaacgcacacccccacgaacacacacacacacacacaca  
chr10 18522096 G 51 @...a.t...aca.ta.a.d...aac,aat,aa.d,.cat,.a.t.t  
chr10 18522097 T 51 @aag,aaaaaga,aa,d,g,ga,cga,ga,aga,.d,aga,aaa,aaaga  
chr10 18522098 G 51 @cccccccccataccacacacca,acaacacacacacaccccccc  
chr10 18522099 T 51 @ggcggggggcggcgcacacgggggggggggggggggggggggg  
chr10 18522100 T 51 @gg,ggggggggggggg.g.ggc,ggggga,gggggg,cgagggga,.d  
chr10 18522101 C 51 @,.aa,d...aga,aa,aaaa,aat,agaaaa,aa,d,agaa,.d,aaa  
chr10 18522102 T 51 @aaaaaaaaaaaaaaaaaaccaaa,aaacaaacaaaaaacaaaaaaaca  
chr10 18522103 T 51 @,.gc.c.g,.ccc.cc,cggg,caa,cacccg,cc.c.gccc,.c.cggc  
chr10 18522104 T 51 @aaa.d.acaa,.d,.d,a,aaa,cag....aa,.d,aa...aa,d,aa,  
chr10 18522105 A 51 @ccggcccccggggggggggggggggg,gggggggggggggggggg  
chr10 18522106 T 51 @gg,.g,ggg,.g.g....g.d.g.c....g.g.g....gg.g....  
chr10 18522107 C 51 @aagagaaaaaggggggggggggggggggggggggggggggggg  
chr10 18522108 C 52 @gggggggggggggggggggggggggggggggggggggggggggg  
chr10 18522109 T 52 @cccccccccccccccccccccccccccccccccccccccccccc  
chr10 18522110 A 52 @tttttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522111 G 52 @cccccccccccccccccccccccccccccccccccccccccccc  
chr10 18522112 G 57 @cccccccccccccccccccccccccccccccccccccccccccc  
chr10 18522113 A 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522114 C 57 @aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa  
chr10 18522115 T 57 @.....cccc  
chr10 18522116 C 57 @gggggggggggggggggggggggggggggggggggggggggggg  
chr10 18522117 A 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522118 G 57 @.....cccc  
chr10 18522119 G 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522120 C 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522121 C 57 @.....aaaa  
chr10 18522122 C 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522123 T 57 @.....ggggg  
chr10 18522124 T 57 @.....aaaa  
chr10 18522125 T 57 @aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa  
chr10 18522126 C 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522127 C 57 @.....aaaa  
chr10 18522128 A 57 @cccccccccccccccccccccccccccccccccccccccccccc  
chr10 18522129 G 57 @tttttttttttttttttttttttttttttttttttttttttttt  
chr10 18522130 A 57 @.....  
chr10 18522131 A 6 @,ggggg  
chr10 18522132 C 6 @,ggggg  
chr10 18522133 A 6 @,.....  
chr10 18522134 G 6 @,cccc  
chr10 18522135 A 6 @,ttttt  
chr10 18522136 A 6 @,cccc  
chr10 18522137 G 6 @,aaaa  
chr10 18522138 A 6 @,ggggg  
chr10 18522139 G 6 @,.....  
chr10 18522140 G 6 @A....
```

Annotating small RNA-seq libraries

- Identify expressed transcripts from trace read alignments to the target genome
- Determine what small RNA components are present
 - Screen for well known structural RNAs (e.g. ribosomal RNA, tRNAs, snoRNAs, etc)
 - Align transcripts to current version of miRbase to identify expressed microRNAs
 - Align transcripts to our own piRNA database built from recently published candidate piRNA sequences
- Set remaining unknown transcript population aside, examine for potentially novel RNAs

High-ranking miRbase alignments

KS35 (Spermatocytes)		KS45 (Round Spermatids)	
microRNA	Depth	microRNA	Depth
mmu-miR-805	1124	mmu-miR-184	5026
mmu-miR-191	472	mmu-miR-28	2799
mmu-miR-298	307	mmu-miR-423-5p	2083
mmu-miR-107	295	mmu-miR-470	494
mmu-miR-99b	290	mmu-miR-191	462
mmu-miR-28	255	mmu-miR-10b	411
mmu-miR-470	247	mmu-miR-34c	342
mmu-miR-151-3p	245	mmu-miR-182	320
mmu-miR-423-5p	220	mmu-miR-16	302
mmu-miR-881	220	mmu-miR-881	272
mmu-miR-184	195	mmu-miR-195	256
mmu-let-7d	108	mmu-miR-465c-5p	255
mmu-miR-34c	107	mmu-miR-743b-3p	161
mmu-miR-103	103	mmu-miR-151-3p	153
mmu-miR-743b-3p	87	mmu-miR-298	132
mmu-miR-202-5p	82	mmu-miR-107	130
mmu-miR-1196	81	mmu-miR-1195	130

KS35 (Spermatocytes)		KS45 (Round Spermatids)	
piRNA	Depth	piRNA	Depth
17446352.13	1364	17446352.13	1581
17446352.12	1322	17446352.12	1419
17446352.11	1201	17446352.11	1313
17446352.1	545	17446352.1	618
17446352.14	249	17446352.14	311
17446352.78	122	17446352.64	136
17446352.93	110	17446352.87	125
17446352.97	109	17446352.97	122
17446352.91	108	17446352.94	120
17446352.67	108	17446352.38	120
17446352.86	107	17446352.96	119
17446352.8	105	17446352.7	118
17446352.54	105	17446352.8	117
17446352.99	105	17446352.86	117
17446352.81	104	17446352.78	117
17446352.82	103	17446352.66	116
17446352.72	99	17446352.89	115

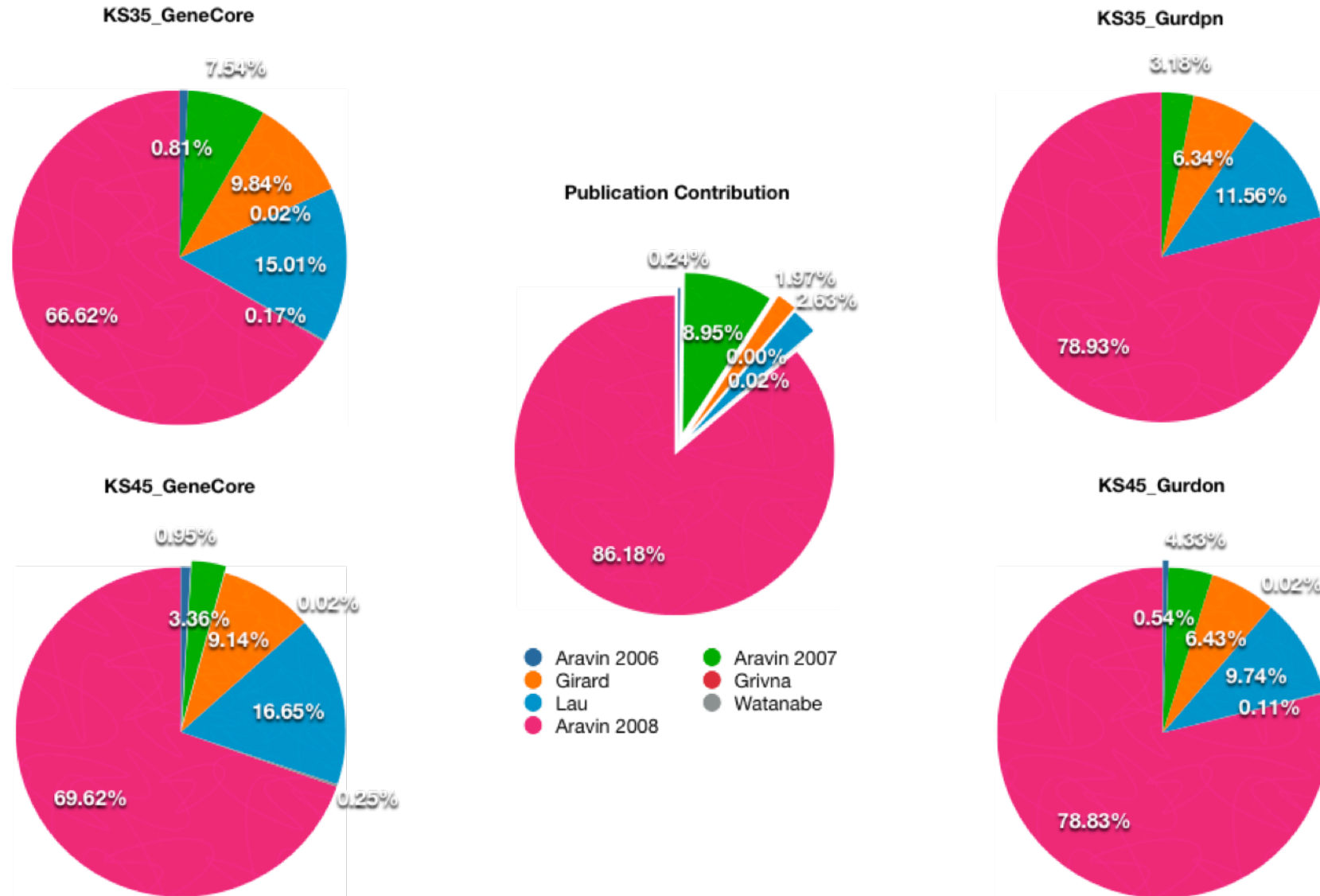
Composition of piRNA Database

Total candidate sequences: 1,524,007

95%, 25nt	Number of piRNAs per publication (Tot=210,576)						
PubmedID	16751777	17446352	16751776	16766680	16778019	16766679	18922463
	Aravin 2006	Aravin 2007	Girard	Grivna	Lau	Watanabe	Aravin 2008
Total/dataset	3,638	136,417	30,024	40	40,102	355	1,313,431
KS45_GeneCore_eland	167	594	1,613	4	2,940	45	12,291
KS35_GeneCore_eland	150	1,390	1,815	3	2,768	31	12,286
KS35_Gurdon_eland	0	808	1,613	0	2,940	0	20,077
KS45_Gurdon_eland	159	1,283	1,907	5	2,888	33	23,371

- ◆ Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006) Characterization of the piRNA complex from rat testes. *Science* 313: 363-367. PMID 16778019. Data came from Table S4. After sorting the table and taking only the uncharacterized sequences there were 40,102 piRNA candidates.
- ◆ Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442: 199-202. PMID 16751776. Data has been deposited into GenBank. There are 30,024 from this study.
- ◆ Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442: 203-207. PMID 16751777. The piRNA candidate sequences are in an Excel table. It's supposed to be one of the files in the supplementary data, but is mislabeled on the website as S3 instead of S4. Removing the known sequences left 3,638 piRNA candidates from this study.
- ◆ Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, et al. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev* 20: 1732-1743. PMID 16766679. Data came from Table S7 (pdf). There are 355 candidate sequences.
- ◆ Grivna ST, Beyret E, Wang Z, Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* 20: 1709-1714. PMID 16766680. Data came from Table SI (pdf). The sequences were only 40 of them.
- ◆ Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* 31: 785-799. PMID: 18922463. Data came from GEO, accession number GSE12757. There are 1,313,431 associated sequences.
- ◆ Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316: 744-747. PMID 17446352. Data came from GEO, accession number GSE7414. There are 136,417 associated sequences.

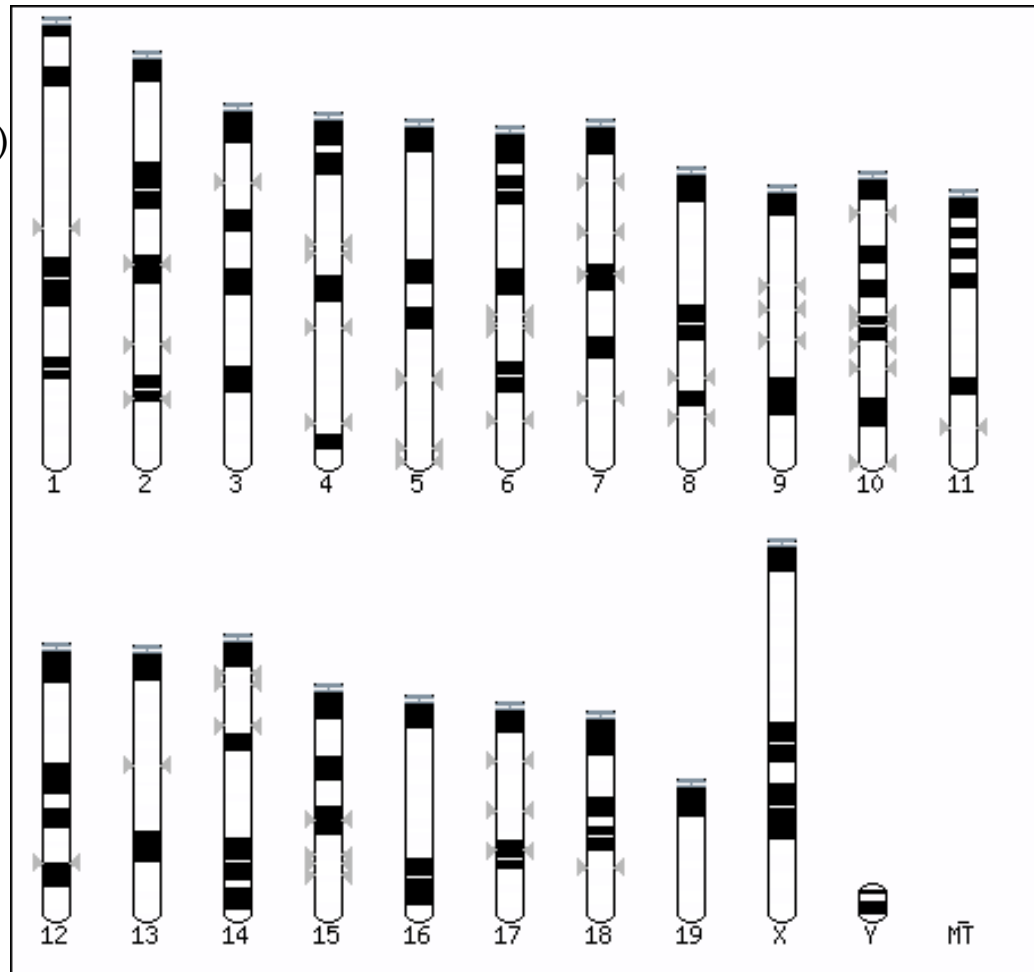
Composition of piRNA Database



Karyogram (I)

Currently annotated piRNA clusters in mouse genome

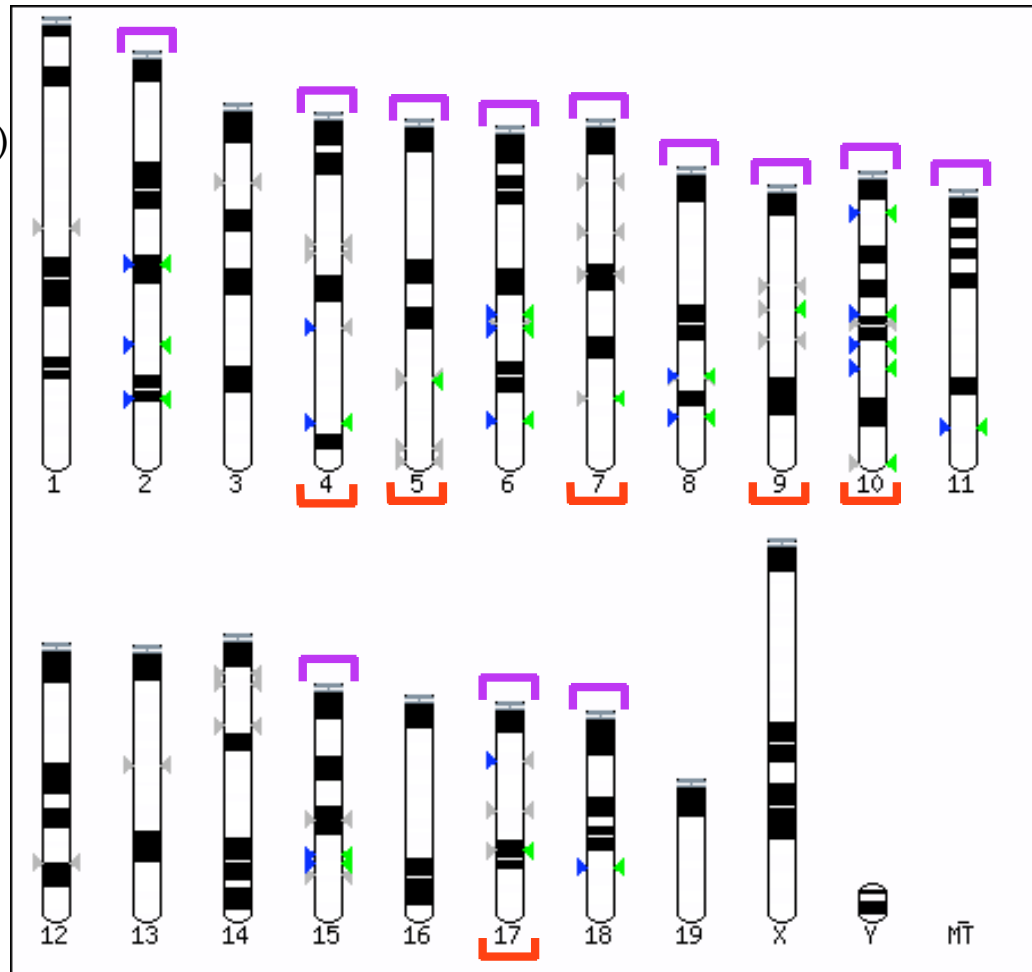
- Left of chromosomes:
KS35 (Spermatocytes)
- Right of chromosomes:
KS45 (Round Spermatids)



Karyogram (2)

Expressed transcripts within piRNA clusters (all levels)

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)

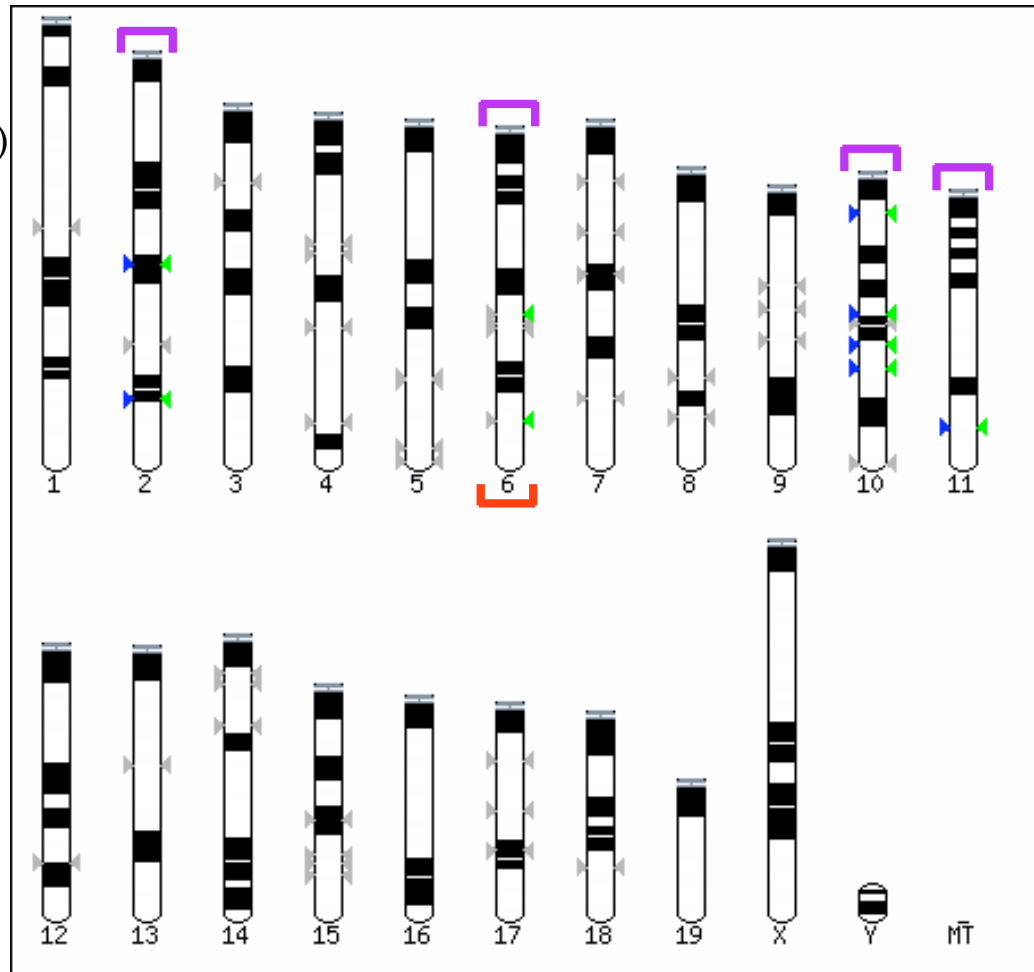


Expression observed
Differential expression observed

Karyogram (3)

Transcript abundance at mid-range levels

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)

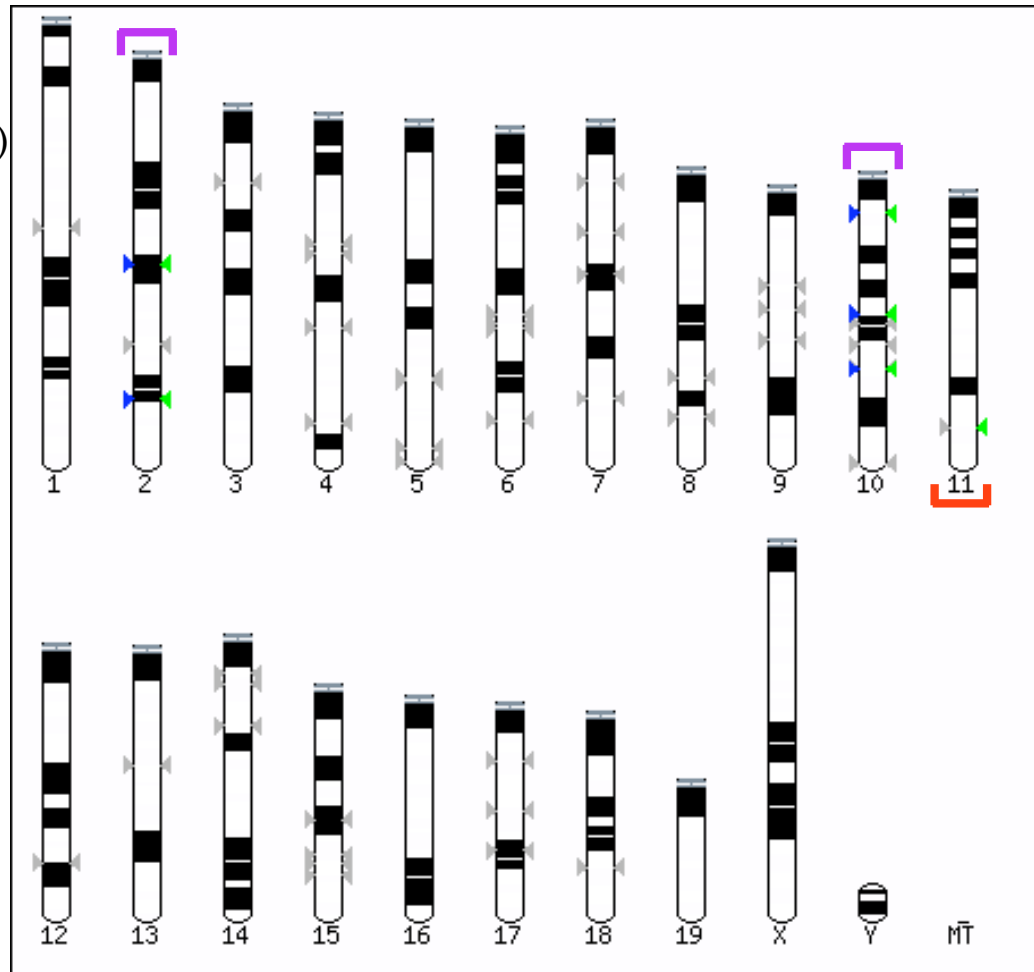


Expression observed
Differential expression observed

Karyogram (3)

Transcript abundance at high levels

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)

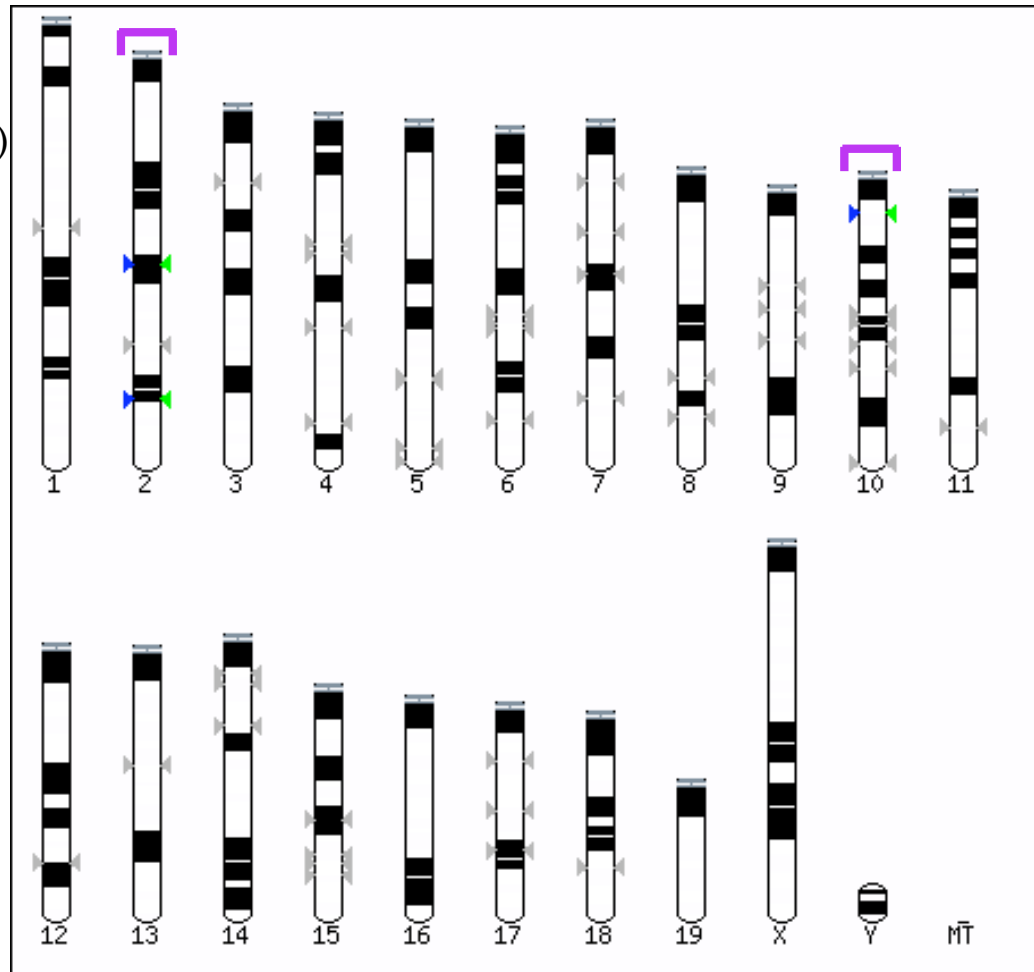


Expression observed
Differential expression observed

Karyogram (4)

Transcript abundance at very high levels

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)

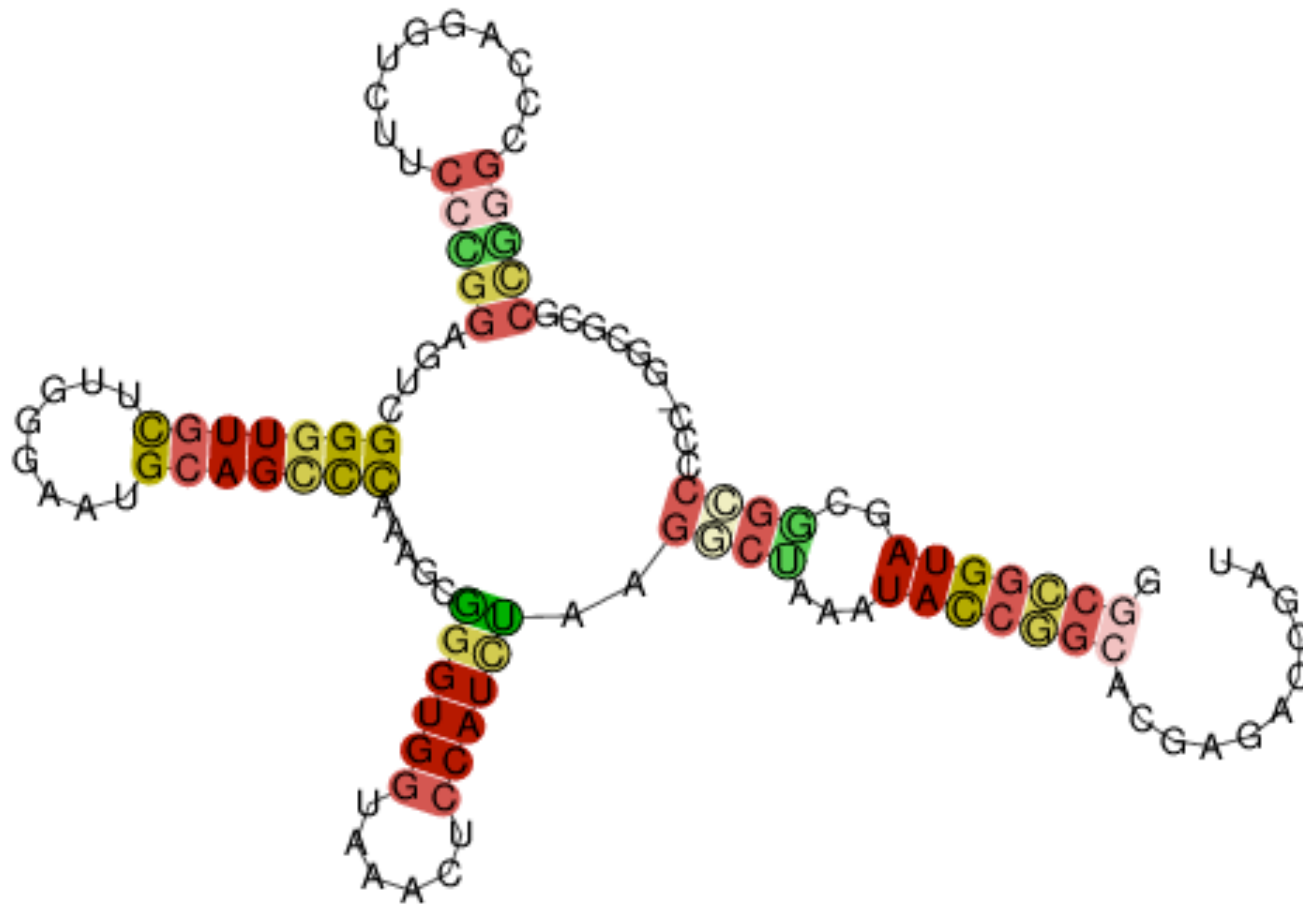


Expression observed
Differential expression observed

Analysis of novel RNA transcripts

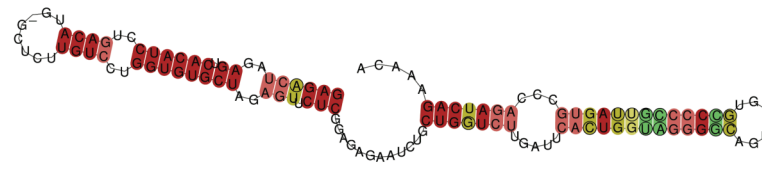
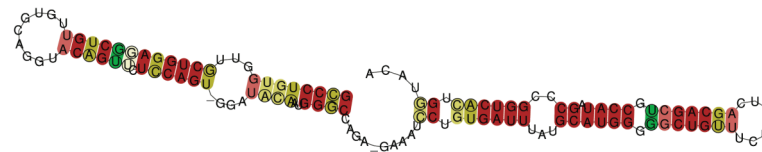
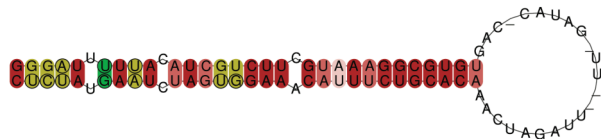
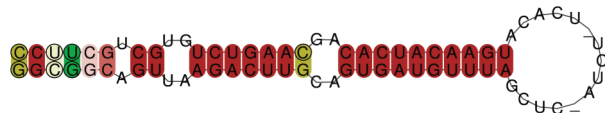
- Transcribed regions fall into several categories
 - Correlate well with annotated (coding) gene loci
 - Correlate with existing non-coding RNAs
 - Novel transcripts
- Novel RNAs
 - To further characterize these, we perform RNA secondary structure prediction on thousands of candidate sequences
 - Look for favorable energy conformations
 - RNAfold (Vienna package), Mfold (Zucker lab)
 - Visualization of putative secondary structures
 - RNAplot (Vienna), StructureLab (Shapiro lab)
 - Homology across multiple species

Prediction of RNA Secondary Structure



Prediction of RNA Secondary Structure

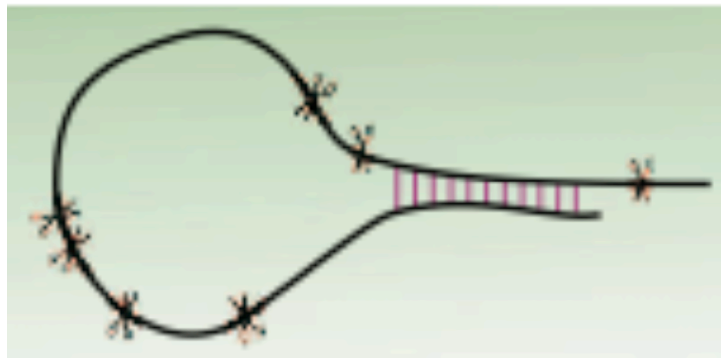
Novel microRNA candidates conserved across species



Stable hairpin consensus structures
 Stem sequences are highly conserved
 Loop sequences are divergent (variable)

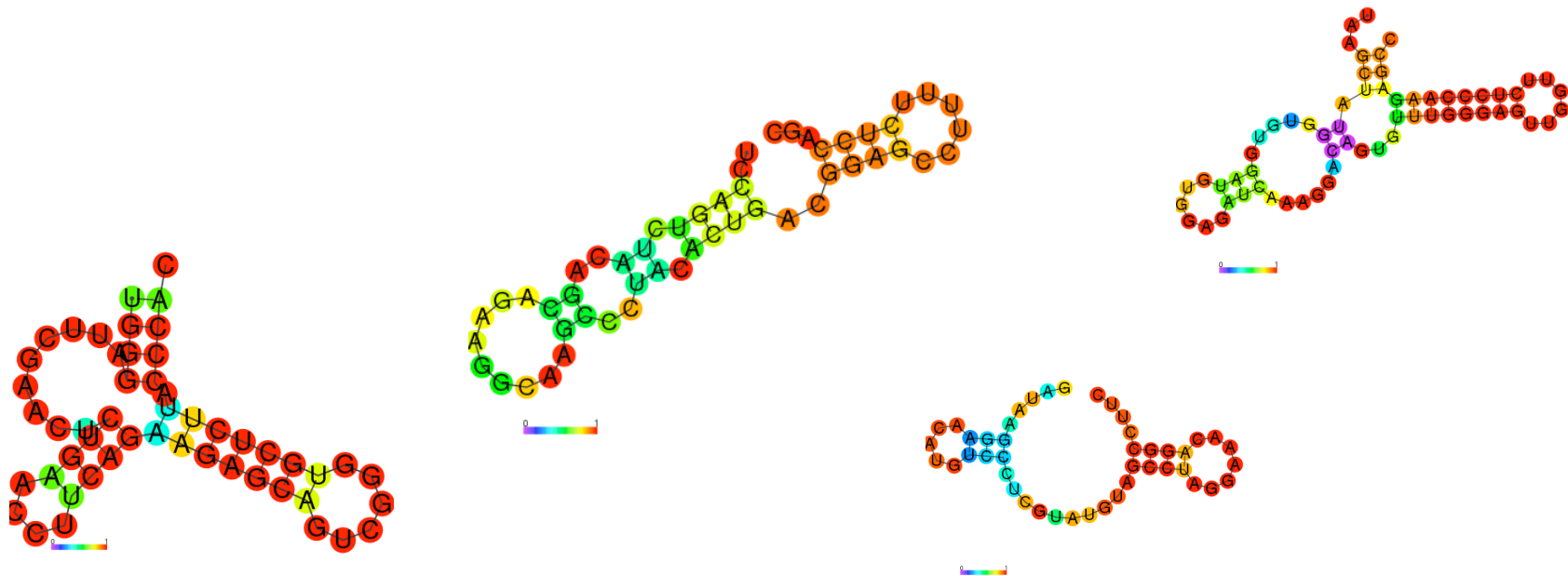
Structural features of piRNAs

- As piRNAs are such a new class of regulatory non-coding RNA, their secondary structural properties are unknown
- Precursor transcripts are processed by a quasi-random mechanism
 - Weak sequence preference near the 5' U



Structural features of piRNAs

- Some structures can be identified based on features typically associated with microRNA hairpins
- It remains to be seen whether these will be characteristic of piRNAs as well



Summary

- Wide variety of RNA sequencing applications
- Library construction protocols differ according to the source material and aims of the experiment
- Open questions about strand specificity, level of coverage required for comprehensive transcriptome analysis
- Single- versus paired-end RNA sequencing
 - As read length increases, sequencing more single-end reads may be more informative